

# York University at TREC 2012: CrowdSourcing Track

Qinmin Hu<sup>1,2</sup>, Zhi Xu<sup>1,3</sup>, Xiangji Huang<sup>1,3</sup>, Zheng Ye<sup>1</sup>

<sup>1</sup> Information Retrieval and Knowledge Management Lab, York University, Toronto, Canada

<sup>2</sup> Department of Computer Science & Engineering, York University, Toronto, Canada

<sup>3</sup> School of Information Technology, York University, Toronto, Canada

vhu@cse.yorku.ca, {zxu, jhuang, yezheng}@yorku.ca

## Abstract

The objective of this work is to address the challenges in managing and analyzing crowdsourcing in the information retrieval field. In particular, we would like to answer the following questions: (1) how to control the quality of the workers when crowdsourcing? (2) How to design the interface such that the workers are willing to participate in and are driven to give useful feedback information? (3) How to make use the crowdsourcing information in the IR systems? The crowdsourcing system called CrowdFlower is employed and four classic information retrieval models are applied in our proposed approaches. Our experimental results show that the IR systems refine the results crowdsourcing by minimizing the manual work and the cost is much less.

## Keywords

CrowdSourcing, Information Retrieval, CrowdFlower, BM25, DFR, Language Model

## 1 Introduction

This is the first year that our York University group participates in the TREC 2012 CrowdSourcing Track. We focus on the text relevance assessing task (TRAT). TRAT is one of the two tasks in the TREC 2012 CrowdSourcing Track, in which the other one is the image relevance assessing task (IRAT). The goal of TRAT is to evaluate approaches to text relevance assessing.

Different with the other tracks such as genomics and blog, we simulate playing the relevance assessing role of NIST in the TREC 8 ad-hoc track [Voorhees and Harman, 1999] with a subset of the TREC 8 topics. 10 topics from TREC 8 ad-hoc are randomly selected for use in the TRAT, which are 411, 416, 417, 420, 427, 432, 438, 445, 446, 447 provided officially by the NIST. The title, description and narrative of a topic define the relevance and non-relevance to itself when a document is judged. The documents are partially selected from the TREC 8 ad-hoc which uses the Text Research Collection Volumes 4 (May 1996) and 5 (April 1997) minus the Congressional Record (CR) [Voorhees and Harman, 1999]. In the TART, there are 18,260 topic-document pairs to be judged under 10 topics.

An assumption is made in the TART that we are required to utilize crowdsourcing to do the relevance assessing, but can use any approach as long as we follow the task's guidelines<sup>1</sup>. Therefore, we propose a crowdsourcing system in Figure with four approaches as the submitted four runs, which obtain the benefits of both crowdsourcing and the traditional information retrieval (IR) models. The differences among these four approaches are that they adopt four relevance feedback methods

---

<sup>1</sup> <http://www.mansci.uwaterloo.ca/msmucker/trec2012TRAT/>

Report Documentation Page			Form Approved OMB No. 0704-0188		
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>NOV 2012</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2012 to 00-00-2012</b>	
4. TITLE AND SUBTITLE <b>York University at TREC 2012: CrowdSourcing Track</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>York University, Information Retrieval and Knowledge Management Lab, Toronto, Canada,</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES <b>Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License</b>					
14. ABSTRACT <b>The objective of this work is to address the challenges in managing and analyzing crowd-sourcing in the information retrieval field. In particular, we would like to answer the following questions: (1) how to control the quality of the workers when crowdsourcing? (2) How to design the interface such that the workers are willing to participate in and are driven to give useful feedback information? (3) How to make use the crowdsourcing information in the IR systems? The crowdsourcing system called CrowdFlower is employed and four classic information retrieval models are applied in our proposed approaches. Our experimental results show that the IR systems refine the results crowdsourcing by minimizing the manual work and the cost is much less.</b>					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>5</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

as interactive feedback, tf-idf feedback, modified pseudo feedback and proximity feedback. The crowdsourcing system we use is CrowdFlower and the traditional IR models are BM25, DFR and language model (LM). There are 47.25 dollars costed in our experiments and in total around 5 hours for both worker training and real jobs. More details are presented in the following sections.

## 2 The Proposed Approaches

Here we present our approaches of (1) how to utilize crowdsourcing; (2) how to apply the IR systems; (3) how to make use of the crowdsourcing information into the IR systems.

The motivation of using the IR systems into the whole crowdsourcing procedure is that the cost will be very high if we manually ask the workers to judge the given 18,260 topic-document pairs directly. Additionally, the accuracy will be less if there are more pairs judged by workers, since the quality control is a big challenge in crowdsourcing. Hence, we refine the number of pairs judged by the workers through retrieving the documents by the IR systems.

As we can see in Figure 1, the IR systems output two rounds results, in which the first round is to obtain the candidate documents for crowdsourcing, and the second round is to provide evidences for the final decisions of the given pairs. A key step of this figure is how to design the crowdsourcing page and get the useful information from the workers. More details will be discussed in the following sections.

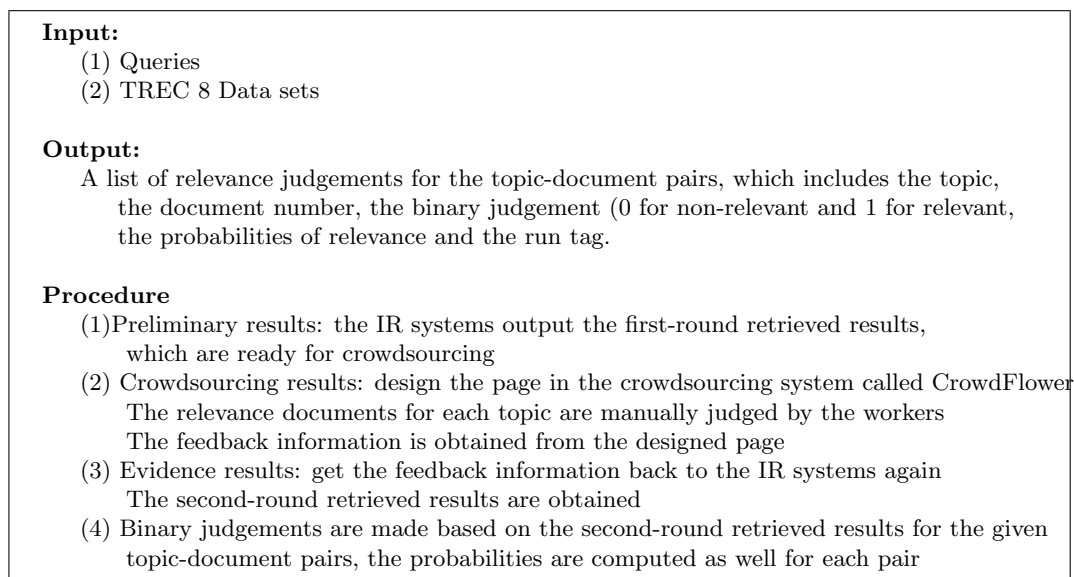


Figure 1: A Whole Procedure of the Proposed Approach

## 3 CrowdFlower

Crowdsourcing is an online practice, which describes the act of outsourcing work to a large group of people of a community as a crowd. It is an open call for contributions from the crowd to complete a task in exchange for social recognition, micro-payments and so on. Nowadays, crowdsourcing has attracted growing attentions as a valuable solution to harness human abilities from a large population of workers [Howe, 2008]. The crowdsourcing of relevance judgements enables the evaluation of the IR systems on the large-scale data sets.

CrowdFlower uses crowdsourcing techniques to provide a wide range of enterprise solutions which process or create large amounts of data. CrowdFlower has over 50 labor channel partners, among them Amazon Mechanical Turk and TrialPay; their network is composed of over 2 million Contributors from all over the world. We present our crowdsourcing stage in Figure 2.

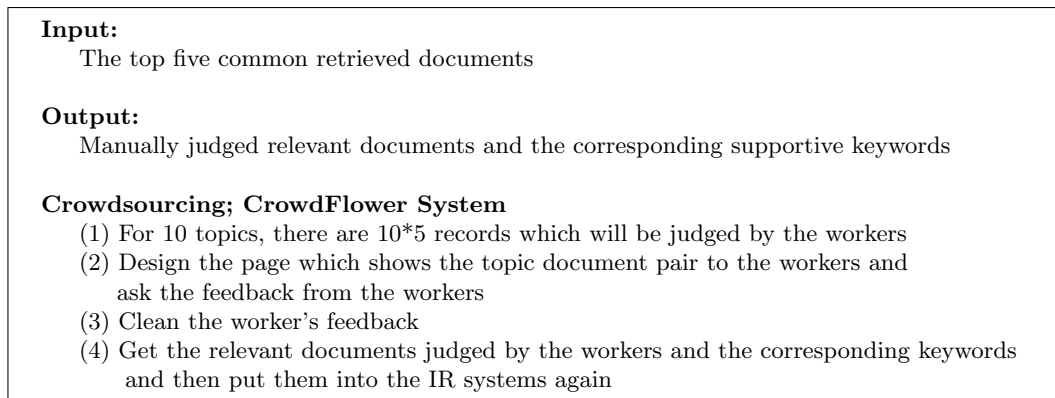


Figure 2: Crowdsourcing Stage

We design our page to catch the feedback information from the workers, including the instructions given for each task, the topic and the document candidate presented for workers' reading. After this general information, the judgements and supported keywords are required for the worker to complete this job.

### 3.1 Quality Control

CrowdFlower stands apart from these individual networks because they offer the quality control, called Gold Standard Data, which has workers perform pre-completed tasks to determine their accuracy and trustworthiness. So we adopt the TREC 7 topics as the training topics to find the valued workers. For each task, we also ask 10 workers to complete the same task as a peer review method. Hence, we present the sample without quality control in Table 1 and the worker information is also listed in Table 2.

unit id	id	judgement	keyword1	keyword2	keywords3	others
195005932	563650952	Relevant	National Assembly	women priests	young women	
195005932	563662790	Relevant	Religious Women LUTHERANS ... BLESSING OF ...	Times Staff Writers	RUSSELL CHANDLER JOHN DART	
195005932	563662882	Relevant	woman	religion	clergy	
195005932	563686956	Relevant	9	7	8	waw
195005932	563690807	Relevant	44	kjlk	h	
195005932	563694956	Irrelevant	f	k	j	l

Table 1: The Sample CrowdSourcing Results

## 4 IR Systems

Here we first present how the IR systems obtain the preliminary results and the evidence results with different IR models in Figure 3.

id	channel	trust	worker id	country	region	city	ip
563650952	amt	1	6012403	USA	IL	Sugar Grove	99.166.149.87
563662790	crowdguru	1	9517560	DEU	1	Stuttgart	85.180.91.110
563662882	getpaid	1	9825820	USA	NY	Hudson Falls	69.205.36.74
563686956	getpaid	0.8571	9742437	ISR	5	Tel Aviv-yafo	77.127.12.98
563690807	amt	0.875	7958171	IND	2	Hyderabad	124.123.79.4
563694956	getpaid	0.2222	9994036	ISR	1	Ashqelon	85.65.227.80

Table 2: Workers' Information

In the preliminary step, the IR systems find the most likely relevant documents as the candidates for the manually judgements in the crowdsourcing stage. In order to improve the relevance possibilities of the retrieved results, four IR models of BM25, DFR, BM25\_DFR and LM are adopted. Note that the candidate for crowdsourcing are the top five common documents in all four retrieved lists.

In the evidence step, the IR systems treat the manually feedback information in four ways: (1) directly make use of the manually judged documents and the keywords provided by the workers in the IR systems; (2) use the manually judged documents and then adopt the proximity feedback method proposed by [Miao et al., 2012] to get the weighted feedback terms; (3) use the manually judged documents and then apply the standard pseudo feedback method [He et al., 2004]; (4) count the TF-IDF of the terms in the manually judged documents and extract the top 20 terms as the feedback terms. These are also our four runs submitted to the track for evaluation.

<p><b>Input:</b></p> <ul style="list-style-type: none"> <li>(1) Queries</li> <li>(2) TREC 8 Data sets</li> </ul> <p><b>Output:</b></p> <ul style="list-style-type: none"> <li>two lists of retrieved documents</li> </ul> <p><b>Preliminary Results:</b></p> <ul style="list-style-type: none"> <li>(1) Customize the queries and the TREC 8 collection into the IR systems</li> <li>(2) Apply an IR model: { Output the retrieval list where documents are searched without no judgements }</li> <li>(3) Four IR models are applied: BM25, DFR, BM25_DFR, LM</li> <li>(4) For each topic { Extract the top five common retrieved documents from four retrieved lists Note that these five retrieved documents have to be retrieved by all the four models}</li> <li>(5) The top five documents for each topic are ready for crowdsourcing</li> </ul> <p><b>Evidence Results:</b></p> <ul style="list-style-type: none"> <li>(1) Make use of the manually judged documents and keywords as the relevance feedback into the IR systems { Output four retrieved lists again under four IR models of BM25, DFR, BM25_DFR and LM}</li> <li>(2) Make use of the manually judged document without keywords { Adopt the proximity feedback method proposed by [Miao et al., 2012] Then get the weighted feedback terms as the relevance feedback Output four retrieved lists again under four IR models of BM25, DFR, BM25_DFR and LM}</li> <li>(3) Make use of the manually judged document without keywords { Adopt the pseudo feedback method [He et al., 2004] Then get the pseudo feedback terms as the relevance feedback Output four retrieved lists again under four IR models of BM25, DFR, BM25_DFR and LM}</li> <li>(4) Make use of the manually judged document without keywords { Calculating the TF-IDF of each term in the relevant documents Extract the top 20 terms as the relevance feedback Output four retrieved lists again under four IR models of BM25, DFR, BM25_DFR and LM}</li> </ul>
---

Figure 3: IR Systems

## 5 Experimental Results

We report our experimental results here. All runs’ binary judgements are evaluated using the LAM measure and treated the adjudicated judgements as truth. The submitted probability of relevance figures are evaluated by the AUC measure and also treated the adjudicated binary judgments as truth.

		run01	run02	run03	run04	Median
LAM	411	0.195	0.238	0.206	0.179	0.15
	416	0.236	0.279	0.187	0.221	0.16
	417	0.277	0.536	0.165	0.199	0.2
	420	0.478	0.47	0.334	0.351	0.17
	427	0.191	0.26	0.184	0.183	0.18
	432	0.468	0.345	0.381	0.364	0.27
	438	0.282	0.422	0.245	0.273	0.26
	445	0.418	0.419	0.192	0.2	0.19
	446	0.499	0.448	0.176	0.205	0.21
LAM ALL	447	0.063	0.077	0.111	0.104	0.08
		0.311	0.349	0.218	0.228	0.187
AUC	411	0.5	0.407	0.474	0.459	0.86
	416	0.471	0.398	0.465	0.48	0.85
	417	0.489	0.5	0.451	0.429	0.75
	420	0.462	0.443	0	0.476	0.71
	427	0.402	0.363	0.464	0.415	0.73
	432	0.47	0.431	0.503	0.454	0.71
	438	0.443	0.448	0.463	0.457	0.78
	445	0	0.403	0.509	0.48	0.83
	446	0.489	0.497	0.491	0.478	0.82
AUC ALL	447	0.498	0.486	0.534	0.523	0.76
		0.467	0.4376	0.479	0.465	0.78

Table 3: Performance of Four Runs Compared to the Official Median Value

## 6 Conclusions and Future Work

Here we present our work in the TART task of the TREC 2012 Crowdsourcing Track. One of our major motivations is to refine the crowdsourcing jobs through adopting the traditional IR systems. The crowdsourcing system called CrowdFlower is employed and four classic information retrieval models are applied in our proposed approaches as BM25, DFR, BM.DFR and LM. Another four feedback methods are also presented as the submitted four runs, where the feedback terms are respectively given by the crowdsourcing workers, by the proximity based feedback method, by the standard pseudo feedback method and TF-IDF.

Our experimental results show that the LAM results is much better than the AUC results. The main reason is that the way that we compute the probabilities is very simple. This is our ongoing work in the near future.

## References

- J. He, M. Li, Z. Li, H. Zhang, H. Tong, and C. Zhang. Pseudo Relevance Feedback Based on Iterative Probabilistic One-Class SVMs in Web Image Retrieval. In *Proceeding of Pacific-Rim Conference on Multimedia*, 2004.
- J. Howe. Crown Publishing Group, New York, NY, USA, 1 edition, 2008.
- Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. Proximity-based rocchio’s model for pseudo relevance. In *SIGIR*, pages 535–544, 2012.
- Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (trec-8). In *TREC*, 1999.